REMARKS

The specification has been amended at page 3, second paragraph, in response to the Examiner's objection to the specification. While the Examiner is technically correct that the specification contains what appears to be embedded and/or other form of browser-executable code, what is in fact described are hypothetical uniform resource locators (URLs) for the purpose of illustrating by example the hierarchical structure of Web pages. To avoid confusion with real URLs, the URLs "www.bank.com/loans", "www.bank.com/loans/auto" and "www.bank.com/loans/homemortgage" have been enclosed in quotes and a parenthetical explanation has been added that these are hypothetical, as opposed to real, URLs for the sake of the example being described. It is believed that this amendment is fully responsive to the Examiner's objection and adds no new matter.

Claims 1 to 6 remain in the application and are re-submitted, without amendment, for reconsideration by the Examiner.

Claims 1 to 6 were rejected under 35 U.S.C. §103(a) as being unpatentable over U.S. Patent No. 6,311,182 to Colbath et al. in view of U.S. Patent No. 5,.819,220 to Sarukkai et al. This rejection is respectfully traversed for the reason that the combination of Colbath et al. and Sarukkai et al. neither shows nor reasonably suggests the claimed invention.

The claimed invention provides an automated method for setting up a Web site with a natural language interface. With reference to Figure 2 of the drawings, in the present method, as claimed, a Web crawler 21, or similar program, creates a hierarchy of topics 22 from the Uniform Resource Locators (URLs) in a Web site (see page 3, lines 5–14, and page 5, lines 19–22, of the present specification). Then, text on each page is analyzed to generate a keyword index 23; each node has an associated collection of selected keywords. These keywords can be n-grams, for example. The use of stochastic n-gram (Markovian) models has a long and successful history in the support of vocabulary applications in speech recognition

systems. Applicants, however, use n-grams in a different way. The logic is as follows. Each topic has a set of n-grams, perhaps sparse, associated with it. Each (sparse) n-gram is connected to a rule in which each term of the n-gram is a term of a rule whose consequent is the topic associated with the n-gram being converted. As used herein, and in the specification, "n-gram" includes sparse n-gram and non-sparse n-gram. The distinction is made on page 3, lines 17–22. Formally, since a sparse n-gram is a set of ordered words (or tokens, etc.) within a window *d*, the traditional notion of an n-gram as a sequence of *n* words, is simply a sparse n-gram with *d=n*; i.e., the length of the sequence with no gaps. In Applicants' usage of n-grams, gaps are allowed between words in their n-grams, hence their n-grams can be sparse or not sparse. As noted in the specification on page 5, lines 7–8 and lines 11–13, the selection criterion can be the chi-square measure, or a statistical test confidence measure. In a final step, a mechanism 25 is specified for associating classification rules to the topic. Classification rules are created from the keywords or n-grams. For example, given the n-gram "need car loan", which is statistically associated with the topic "car_loan", the rule "need & car & loan → car_loan" can be produced. This rule can be associated with topics relating to cars or loans.

In making the rejection, the Examiner alleges that Colbath et al. teach "An automated method for setting up a natural language interface in a Web site", but as will be demonstrated below, this is not true. The Examiner further alleges that Colbath et al. teaches the steps of "defining" and "generating" as recited in independent claim 1, but again as will be demonstrated below, this is also not true. The Examiner states that "Colbath does not explicitly teach, 'for each topic in the hierarchy, a set of n-grams to a topic in the topic hierarchy which set of n-grams is distinctive to the topic and wherein the n-grams maybe sparse or non-sparse n-grams" (emphasis added). It is noted here that Colbath et al. neither explicitly nor implicitly teach this feature. The Examiner relies on Sarukkai et al. for a teaching of this feature, citing column 7, line 27, to column 8, line 11, and column 10, lines 16 to 24, of Sarukkai et al. However, Sarukkai et al. neither shows nor suggests

this feature. In fact, the notion of non-sparse n-grams is unique to the claimed invention and, furthermore, the application of n-grams as described in the subject application is unique to the claimed invention.

Considering first, the patent to Colbath et al., Colbath et al. teach a very different technology from that of the claimed invention; specifically, a voice-activated Web browser. In Colbath et al., voice signals are recognized and converted into words. These words are used to form a search string, and a search is then performed, for example, on the Internet or on a Web site. The search is performed over a preselected collection of areas of interest. Colbath et al. further disclose methods for searching when the search terms do not match with any preselected areas of interest.

Colbath et al. is very different from the claimed invention for several reasons. First, the claimed invention is directed to a method for setting up a Web site query interface, and Colbath et al., by contrast, is directed towards searching based on voice commands. Colbath et al. do not teach setting up a Web query interface, as alleged by the Examiner. Second, as recognized by the Examiner, Colbath et al. do not teach the step of, for each Web site topic, associating a set of n-grams to the topic, which are distinctive of that topic, as recited in the third step of claim 1. In the preferred embodiment, these sets of n-grams are converted to classification rules, and claim 6, dependent on claim 1, recites this step.

Colbath et al. do not teach or suggest an automatic method for setting up a Web query interface, as alleged by the Examiner. In fact, Colbath et al. is completely lacking any suggestion to set up a query interface. Instead, Colbath et al. teaches only methods for conducting Web searches using voice commands.

By comparison, independent claim 1 and dependent claim 3 are directed to "setting up a natural language interface in a Web site". Setting up a natural language interface according to the present invention requires that documents on a Web site are classified, and requires that a keyword index is created for documents in the Web site. This allows a person creating the natural language interface to do so efficiently and easily. The natural language interface allows a search engine to

find documents on a Web site set up according to the invention. Colbath et al., do not teach how to create or set up a natural language interface, but instead teach how to perform a search using voice commands. Setting up a natural language interface and performing a search are two entirely different and distinct functions. Setting up a natural language interface allows a search program to search a Web site according to a query protocol (possibly specified by the interface), and performing a search finds documents of interest. Hence, the teachings of Colbath et al. are not really applicable to the claimed invention.

Specifically, because Colbath et al. do not teach setting up a natural language interface, and instead teach performing a search, they necessarily lack, contrary to the Examiner's allegation, the essential step of "generating a keyword index for those documents", as recited in claim 1. The Examiner argues that Colbath et al. teach this limitation in col. 3, lines 1–12. However, in this passage, Colbath et al. explain something quite different; specifically, that it is the "most probable word strings" of the *input speech* that are searched for. By comparison, in the claimed invention, the above-referenced limitation requires that a keyword index is *created for a collection of documents* so that the documents can be searched more effectively. The keyword index of the present invention allows a search engine to find documents; *the keyword index is not searched for*, as required by Colbath et al. Instead, the keyword index of the present invention *represents a field searched in*. The Examiner has confused the search terms with the search field in the Colbath et al. reference. Hence, the teachings of Colbath et al. do not include or suggest generating a keyword index as in the present invention.

Also, as noted above, Colbath et al. does not teach a mechanism for associating a rule to a topic, as required by claim 1. The claimed invention, and in particular the third element of claim 1, is not concerned with speech recognition (although it may be compatible with speech recognition). The third element of claim 1 requires that each topic in the topic hierarchy is associated with a set of n-grams which are distinctive of that topic, so that searches can be performed.

Regarding claim 3, the Examiner argues that Colbath et al. teach a keyword index, and that reviewing the keyword index is also taught by Colbath et al. However, Colbath et al. do not teach a keyword index according to the present invention. Col. 2, lines 20–35, of Colbath et al., identified by the Examiner with reference to claim 3, teaches that key words are searched for by providing them to a search engine. Col. 2, lines 20–35, does not teach a keyword index as in the present invention, wherein the keyword index is created from Web pages and is a field searched in. Hence, Colbath et al. do not meet the limitations of claim 3.

Regarding claim 4, Colbath et al. do not teach "creating rules from the sparse n-grams, wherein each topic has associated rules that are used to decide if a new input document or query references the topic". This is because Colbath et al. do not teach a natural language interface, and Colbath et al. do not teach that topics have associated rules. Colbath et al. teach only a voice activated search or Web browser, as explained above. The above-quoted limitation from claim 4 requires that Web pages or documents be classified into a topic hierarchy so that they may be searched according to the present invention. Colbath et al. do not teach setting up topics or classifying data so that it can be searched, and hence do not meet this limitation of claim 4.

Sarukkai et al. do teach the use of n-gram language models. However, the teachings of Sarukkai et al. are not applicable to the claimed invention because they are not directed toward the set-up of a natural language interface. Sarukkai et al. instead teach methods for dynamically altering language models according to word sets in the documents searched. In other words, the language model is adjusted in response to documents found in a search. The n-grams used by Sarukkai et al. are used for speech recognition, as known in the art. For example, Sarukkai et al. teach smoothing or re-estimating "n-gram *language model scores*..." (col. 9, lines 20–21, emphasis added), thereby implying that the n-grams are used for speech recognition. N-grams are extremely well known in the art of speech recognition. By comparison, the n-grams employed in the present invention are created from documents to be searched, and the n-grams are stored

as an index for searching. Hence, the n-grams in the present invention are used for very different purposes compared to the n-grams of Sarukkai et al. Consequently, the n-grams of Sarukkai et al. cannot reasonably be combined with Colbath et al. to meet the limitations of claims 2 or 4, as the Examiner argues.

Much of the confusion on the part of the examiner comes from two sources: (1) the failure to distinguish the field of speech/voice recognition and generation/synthesis from text-based natural language processing, e.g, as ubiquitous in search applications and (2) failure to distinguish a method for setting up a system, as in the current invention, from the systems themselves. Beyond that, in the two patents referred to and the other references, there is no mention of automated methods for setting up any system let alone a Web-based natural language interface.

To review the claimed invention, the basic set up is the following:

1. The system implicit in the invention, to which the automated set up methods pertain, requires a taxonomy of topics for a collection of documents, assumed to be associated with URLs, and a set of classification rules for each topic. The classification rules are used to classify user queries into topics as described in the now issued patent, cited as patent application Serial No. 09/570,788 in the cross-reference to related applications on page 1 of the specification.
2. The claimed invention specifies how to induce a taxonomy from a set of URLs and their associated documents and then a set of classification rules for the nodes in the taxonomy.
3. The method consists of (i) crawling a particular Web site, producing a set of Web pages (the documents to be associated with a taxonomy); (ii) using the structure of the URLs as the structure of the hierarchy; (iii) extracting from individual documents and from groups of documents, so-called sparse n-grams, each of which is characteristic of a document or group of documents, where each group is associated with a node in the taxonomy; (iv) determining which phrases, whether sparse or not, are characteristic of

characteristic of the document or group of documents by some statistical technique for identifying salient collocations; and (v) converting the so-called sparse n-grams to classification rules for use in a classifier as described in patent application Serial No. 09/570,788.

Note that the term "sparse n-gram" as defined and used in the disclosed and claimed invention, are sequences of tokens or words from the text where the tokens or words may or may not have other words between them. Perhaps the term "sparse n-gram" has confused the Examiner into thinking that the n-grams as used in art of speech/voice recognition is relevant to the claimed invention. However, both the specification as filed and the foregoing explanation have made clear that the claimed invention is using the concept of n-grams in a different way than used in the art of speech/voice recognition. All that is meant is the more generic notion of a set (or sequence) of not necessarily adjacent tokens or words in the text. So for instance, in a document about Mortgage Loan Applications, one would presumably identify the phrase "Mortgage Loan" or even the noncontiguous phrase "Mortgage Application" as characteristic of the document. An alternative description would be "sparse phrases", and if this helps the Examiner to better understand the disclosed and claimed invention, he is invited to substitute that description for the term "sparse n-gram". Note also that there are two subcases of determining distinctive collocations (sparse phrases, sparse n-grams): those distinctive of a single document and those distinctive of a group of documents. Many methods for doing this are well understood in the art and which is used is not material to the general idea of the disclosed and claimed invention.

While at least one of the cited references mentions crawling the Web as part of a search engine, the use to which the crawling of the Web is put is entirely different. None of the literature or patents cited touch on the items above. Specifically, none of them mention using a taxonomy of topics let alone inducing a taxonomy. As the current invention is not about the specific use of the taxonomy or classification rules (this is covered in patent application Serial No. 09/570,788) and none of the cited references or patents mention this, it can be seen that they do

not say anything relevant about this key part of the invention.

None of the literature or patents cited mention using so-called sparse n-grams in the manner used in the current invention, namely, in conjunction with documents and groups of documents associated with nodes or topics in an (induced) hierarchy to identify collocations or phrases that are characteristic of the associated document or group of documents.

None of the literature or patents cited mention converting sparse n-grams or collocations into classification rules, whose use is described in the context of a classification-based natural language interface for the Web in patent application Serial No. 09/570,788.

It follows from this that none of the cited literature or patents deal in any way with the combination of these methods nor is such combination implicit in the cited literature or patents singly or in combination. It certainly cannot be reasonably maintained, when this is understood, that the claimed invention is anticipated or made obvious by the references or combination of references. Nor can it be reasonably maintained that the claimed invention is an obvious extension or alteration of what is taught in the references.

Briefly summarizing, Colbath et al. deal with a speech or voice interface that involves simple key word matching against a database of topics or microdomains and associated predefined keywords or phrases. Colbath does not discuss setting up a taxonomy or hierarchy of topics, let alone one induced from a set of URLs. Nor do Colbath et al. discuss or mention building a set of classification rules from the content associate with a taxonomy or topic hierarchy induced from a set of documents associated with URLs. Colbath et al. do not discuss how one identifies the topics or micro-domains nor how to establish the predefined phrases. In contrast, the claimed invention deals exclusively with a method for inducing or automatically setting up a taxonomy of topics (Sarukkai et al. are silent on the matter of hierarchically structured taxonomies) and with automatically inducing phrases or sparse n-grams distinctive of documents or groups of document associated with nodes or topics in the automatically induced

taxonomy. So, the claimed invention and Colbath et al. treat entirely different topics.

Sarukkai et al. deal with a voice activated browser. In large part, Sarukkai et al. deal with how to overcome problems with speech recognition algorithms when there are words that are "out of vocabulary". Instead of employing rewriting style grammar, which is very rigid, Sarukkai et al. employ n-grams as they do not impose strict word order constraints. But n-grams also have the problem that they are statically trained on a given corpora and the Web will always have many words not in the training corpus, which means the speech recognition system. The claimed invention deals with dynamically altering scores of the statistical language model and acoustic model used in speech recognition systems. Sarukkai et al. simply do not deal with any of the topics addressed in the disclosed and claimed invention. The common use of the term n-gram, which at a technical level are quite distinct, as for Sarukkai et al., "n-gram" means a sequence of tokens that are assigned probabilities with the context of a speech recognition system language model, is irrelevant to the claimed invention. Many systems use common technologies, but even here the details of usage are very different. One cannot reasonably maintain that Sarukkai et al. anticipates or teaches any features the claimed invention. Nor can anyone maintain with reason that the combination of Sarukkai et al. and Colbath et al. provide what is claimed as neither one treats any of the key items listed above.

The newly cited literature has been reviewed, but none address what the claimed invention does, nor do they collectively.

The reference to www.w3.org/TR/vocie-grammar, "Grammar representation requirements for voice markup language" is a working group paper proposing standards for developing grammars for use in the speech recognition part of an interactive "voice browsers" (cf. "speech recognition grammar specification language that will be useful across a variety of speech platforms used in the context of a dialog and synthesis markup environment" from, page 2 of the document (O. Introduction). The goal is to standardize grammars developed by

different people to enable reuse. As such, the document deals with the format and conventions for the development of such grammars. For instance, Section 2 (Large Vocabulary and Dictation) deals with requirements and specifications for the definition of large vocabularies for dictation applications. Section 8 (Grammar Specification Language) deals with the requirement that grammars be understandable, support extensions, etc. Moreover, in the practice, such grammars are manually developed. This reference does not address any of the problems dealt with in the disclosed and claimed invention. The claimed invention deals with a method for automatically setting up a Web-based natural language interface, and it is made clear from the invention description, the input is text, not speech.

The reference to "A Natural Language Processing Based Internet Agent" deals with an artificial intelligence (AI) based system for understanding a user's query, where the system interacts with a meta Internet search engine (cf. Abstract). The architecture (Figure 1 NIAGENT architecture) shows several components: (1) Niagent, (2) Paragent, (3) MIT Chopper, (4) Metacrawler, and (5) Spider. The approach is to build a system as a "society of interacting agents" (Section 2.1). At a general level, this is a complex, manually developed system. The system uses the MIT Chopper to understand a user's natural language query. The basic idea is to use a meta-search engine to return a large number of documents in response to a query (high recall) and then to select from this large pool of documents, relevant ones by further, more sophisticated natural language processing (high precision). A document is deemed relevant if all the phrases in a query match. But since the same phrases can appear with different relationships among them, simply returning documents based on phrasal match is too expansive. MIT Chopper is a parser, analyzing a sentence into its constituent parts. The rules of the system are hand coded (manually developed). A query is first sent to MIT Chopper to return the relevant phrases used in the meta-search (Section 2.1). The next step is to determine whether the matched phrases likely occur in the right relationships among the matched phrases in the return documents. This is done by the PARAGENT (2.3) – the idea is to divide documents into so-called logical

segments (usually paragraphs), and then the authors claim that PARAGENT determines the relationship between any matched phrases in the same logical segment. The paper does not describe how the system determines the relationship between matched phrases or how it is determined that the relationship is the same as the one in the user query. The description is in this sense incomplete. In summary, the reference describes a two stage system for general search of the Web: the first step is meant to achieve high recall. But in any event, the reference does not deal with the topic of automatically setting up a Web-based natural language system in the sense of the current invention, that is, it does not address any of the topics listed above. Moreover all the methods and techniques are different from those of the claimed invention.

The reference "Integrating Web Resources and lexicons into a Natural Language Query System" deals with a use of a natural language parser for Web question and answering systems. The system is manually developed. There is no discussion of automatically setting up the system. The system uses completely different techniques, algorithms, technologies from the current invention. There is simply no discussion of the items listed above that are key to the claimed invention, e.g., the system does not use a topic hierarchy nor classification methods.

The reference "Intelligent Web Representations" describes a complicated system based on typical AI techniques: natural language parser, knowledge representation. The system is manually developed. It does not use a topic taxonomy to classify Web documents, let alone discuss how to automatically set up such a taxonomy or automatically induce the classification rules, which is the key point of the claimed invention. This reference is simply not relevant to the disclosed and claimed invention.

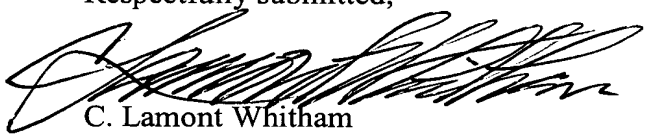
In view of the foregoing, it is respectfully requested that the application be reconsidered, that claims 1 to 6 be allowed, and that the application be passed to issue.

Should the Examiner find the application to be other than in condition for

allowance, the Examiner is requested to contact the undersigned at the local telephone number listed below to discuss any other changes deemed necessary in a telephonic or personal interview.

A provisional petition is hereby made for any extension of time necessary for the continued pendency during the life of this application. Please charge any fees for such provisional petition and any deficiencies in fees and credit any overpayment of fees to Attorney's Deposit Account No. 50-2041.

Respectfully submitted,



C. Lamont Whitham
Reg. No. 22,424

Whitham, Curtis & Christofferson, P.C.
11491 Sunset Hills Road, Suite 340
Reston, VA 20190
Tel. (703) 787-9400
Fax. (703) 787-7557